University of Málaga

Health Engineering

Laboratory Task

Nonparametric Models

Author

Alejandro Dominguez Recio

Course

Intelligent Systems

Teachers

Enrique Domínguez Merino

Jesús de Benito Picazo

Introduction

In this practice we going to evaluate three diferents dataset with the Ibk classifier which is an implementation of the k-nearest neighbours classifier. To evaluate which dataset is more optimal with this classifier we going to describe the performance measures seen in class like confusion matrix, acurracy, precision, fallout, recall, F-measure and area under ROC curve. In order to improve its performance, we will modify the parameters such as the number of closest neighbors and the distance function. Also we going to detail the characteristics of the different datasets.

How are we going to do it?

To do the different studies we will have the following structure in each of then.

- ➔ **Dataset context**

In this part we going to write a little introduction of the dataset context commenting on data type and prediction target.

- ➔ **Number of classes**

We going to describe the number of classes as if it is binary o multiclass. To do that we going to observe the atributtes weka panel and we going to select the class attribute and depending on the number of values that it takes, we will determine if it is binary or multiclass.

- ➔ **Number of attributes**

For the number of attributes we going to inspect the weka attributes panel or open the file in text mode and inspect the characteristics.

- ➔ **Number of samples or instances**

To know the number of samples we can proceed as in the previous section by inspecting the weka panels or opening the document in text mode and seeing its characteristics.

➔ **Performance metric values by configuration**

Accuracy

We can obtain the acurracy in two ways, one of them is directly from the classifier out and the another one is by calculating the number of correctly predicted examples divided by the total number of examples.

Precision

This measurement shows the positive predictive value, higher is bettter. We are going to obtain the necessary values from the confusion matrix and perform the following calculation TP/(TP+FP).

Fallout

This measurement shows the false positive rate, lower is better. We are going to obtain the necessary values from the confusion matrix and perform the following calculation FP/(FP+TN).

Recall

This measurement shows the true positive rate, higher is better. We are going to obtain the necessary values from the confusion matrix and perform the following calculation TP/(TP+FN).

F-measure

This measurement provides a single score that balances both the concers of precision and recall in one number, higher is better. We are going to obtain the necessary values of the previous measurements, precision and recall, and perform the following calculation 2*(Precision*Recall)/(Precision+Recall).

➔ **ROC curve and the area under the curve**

We are going to obtain this measurement by visualizing the threshold curve in Weka. The area under the ROC curve is a number in the interval [0,1], a higher value is better. This measure show the trade-off between the ratios of false positives and false negatives. This measurement is very useful when we faced with unbalanced data.

*The previous points will be repeated for each dataset

➔ **Evaluation of the results**

In this section we are goint to compare the results of the different performance measures.

➔ **Differences between datasets**

Here we are going to describe the main differences in the data of the different datasets.

➔ **Last conclusions**

Finally, we are going explain possible reasons why some datasets perform better with the Naive Bayes classifier than others. We will support our conclusions on the performance measures taken and on the characteristics of the datasets

BREAST CANCER DATASET

➔ Introduction

The main idea of this classification problem is given a dataset about patients with a series of characteristics which determine if they have recurrence events or no recurrence events of breast cancer train a Naive Bayes classification model and evaluate its performance.

In this dataset we have 286 instances in total of which 201 of one class and 85 instances of another class. Each instance has ten attributes, one of them is the class attribute. Because of we have only two classes we are faced with a binary classification problem.

➔ Number of classes

In this case our class attribute can take only **two *values***, *recurrence events* and *no recurrence events*. Because of that we are faced to a ***binary classification problem***.

➔ Number of attributes

In this case we have **ten attributes** which are *class, age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad* and *irradiat.*

➔ Number of samples

In this case in particular the number of samples or instances is **286.**

➔ **Performance metric values by configuration**

Configurations

Configuration	KNN	Distance Function
C1	1	Euclidean
C2	1	Manhattan
C3	1	Minkowski
C4	4	Euclidean
C5	4	Manhattan
C6	4	Minkowski
C7	9	Euclidean
C8	9	Manhattan
C9	9	Minkowski
C10	6	Minkowski
C11	12	Minkowski
C12	15	Minkowski

Results

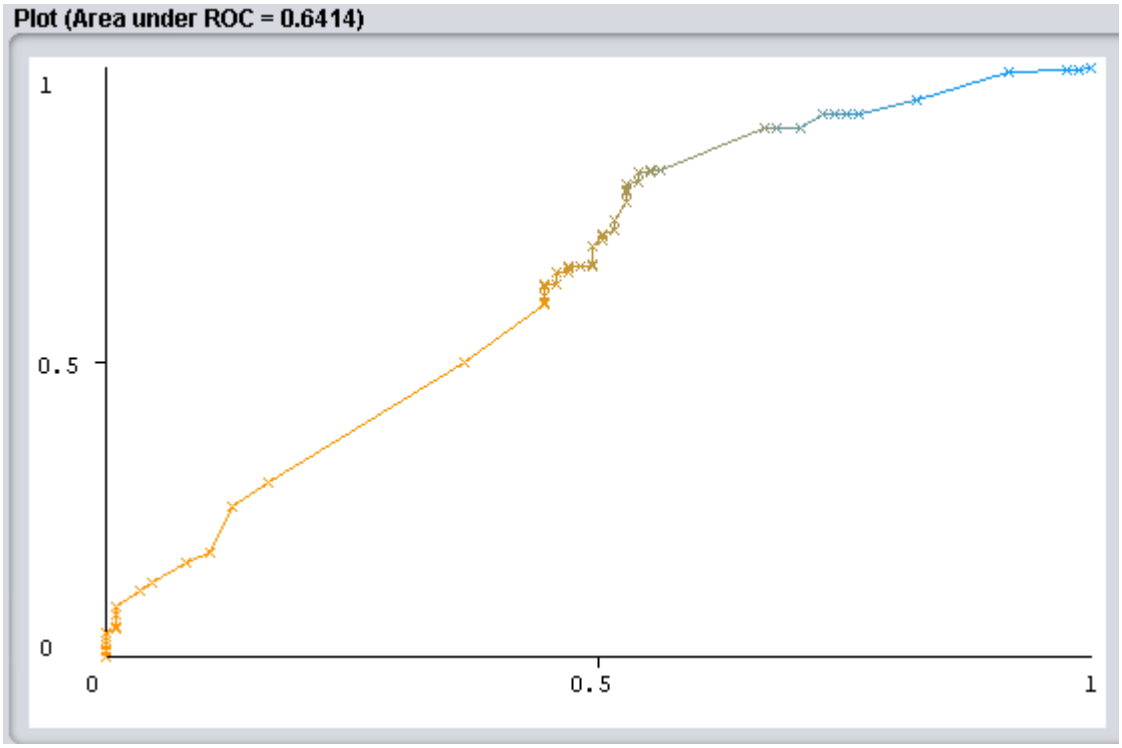
Dataset	Correctly Classified%	TP rate	FP rate	Precision	Recall	Fmeasure	ROC
C1	72,7273	0,896	0,671	0,759	0,896	0,822	0,641
C2	74,8252	0,975	0,788	0,745	0,975	0,845	0,670
C3	74,1259	0,970	0,800	0,741	0,970	0,841	0,676
C4	72,7273	0,896	0,671	0,759	0,896	0,822	0,641
C5	74,8252	0,975	0,788	0,745	0,975	0,845	0,670
C6	74,1259	0,970	0,800	0,741	0,970	0,841	0,676
C7	72,7273	0,896	0,671	0,759	0,896	0,822	0,641
C8	74,8252	0,975	0,788	0,745	0,975	0,845	0,670
C9	74,1259	0,970	0,800	0,741	0,970	0,841	0,676
C10	73,776	0,970	0,812	0,739	0,970	0,839	0,670
C11	73,0769	0,965	0,824	0,735	0,965	0,834	0,685
C12	73,7762	0,985	0,847	0,733	0,985	0,841	0,676

Section conclusions

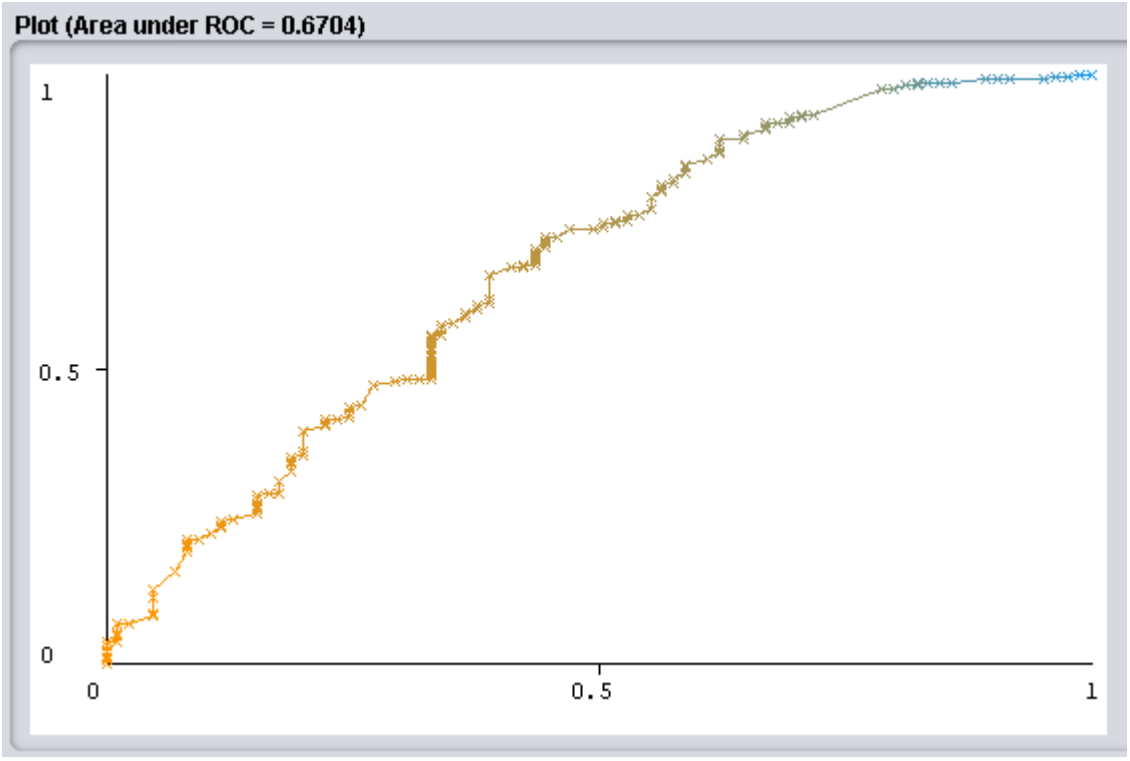
In the first place, the selection of the values of k has the objective of seeing how the model behaves according to its size and its effect on noise as in even and odd values and its effect on class ties. We also change the distance function.

As we can, configuration 1 is the one that gives us the worst result, stating that a value of k = 1 creates a lot of noise in the classification. The best results in correctly classified instances can be found with k = 4 although the value of K = 12 gives us the best ROC area. By way of conclusion, we can say that this dataset behaves in a similar way in the tested configurations, perhaps because its instances are clearly differences and contribute little noise to the classification.

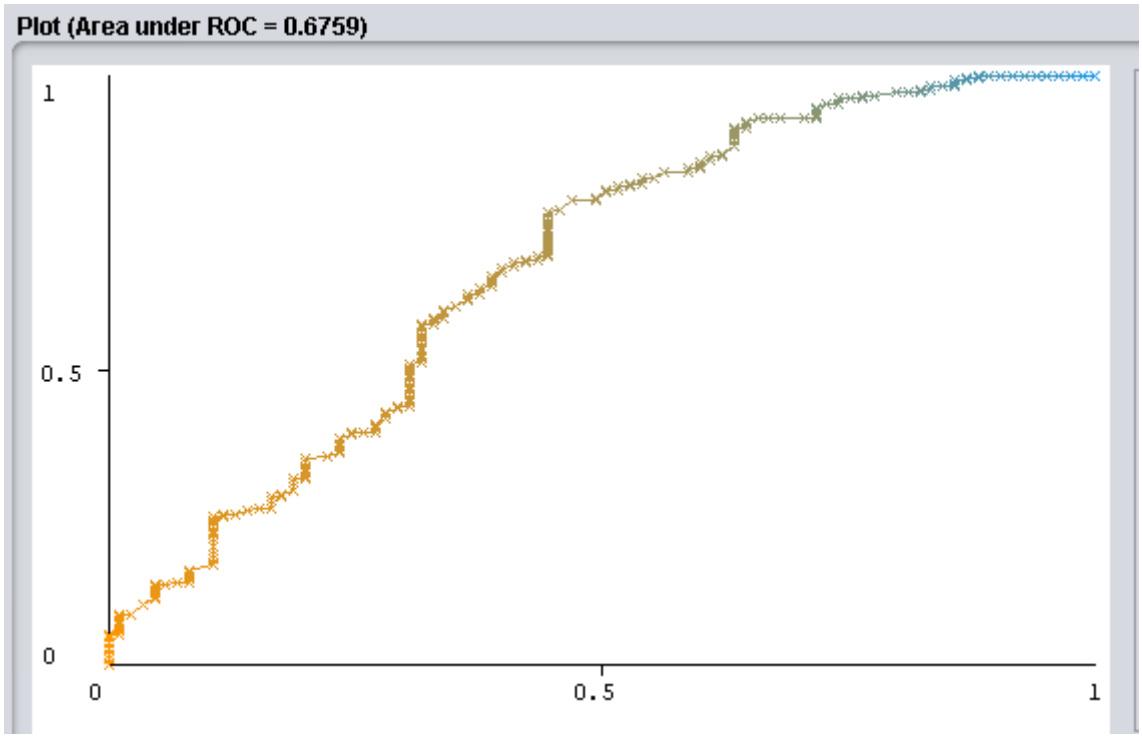
➔ **ROC curve C1**



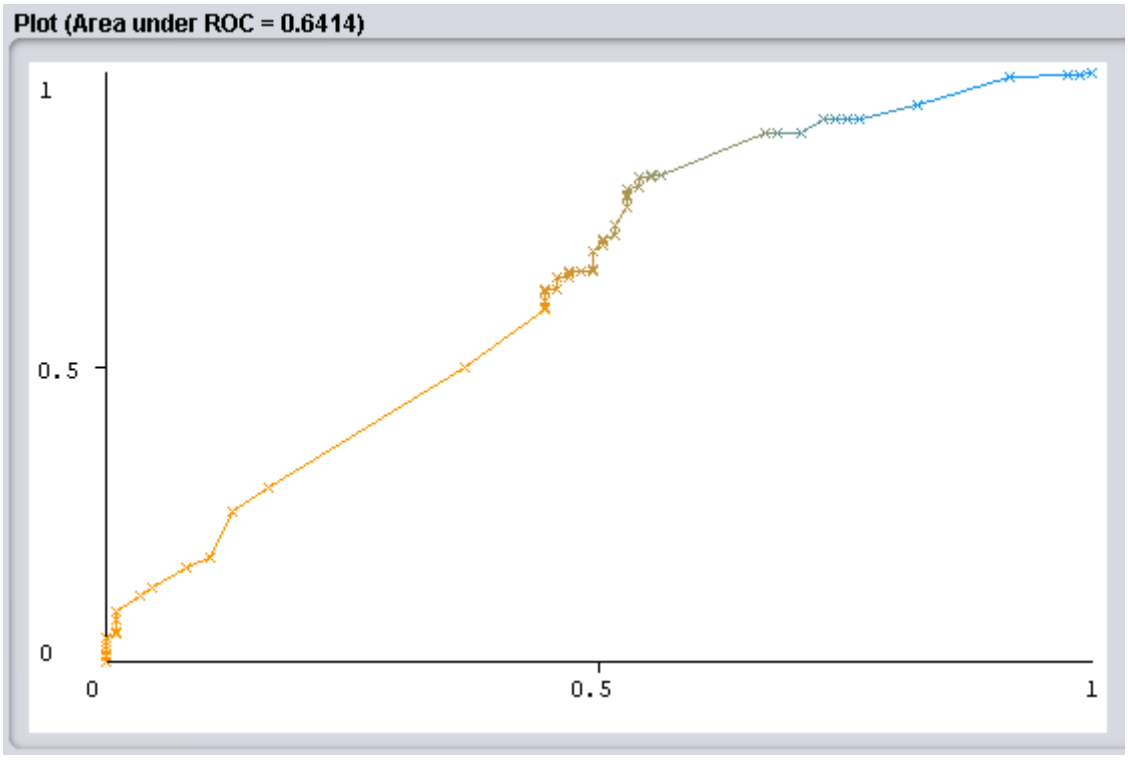
➔ **ROC curve C2**



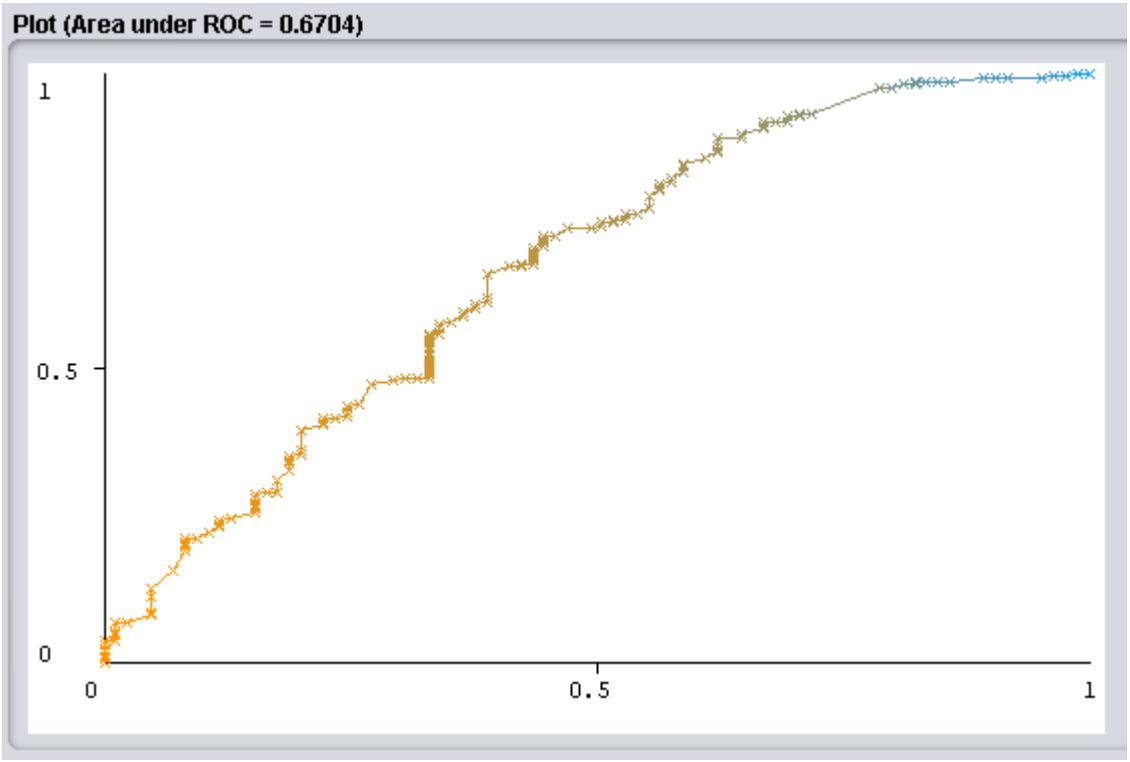
➔ ROC curve C3



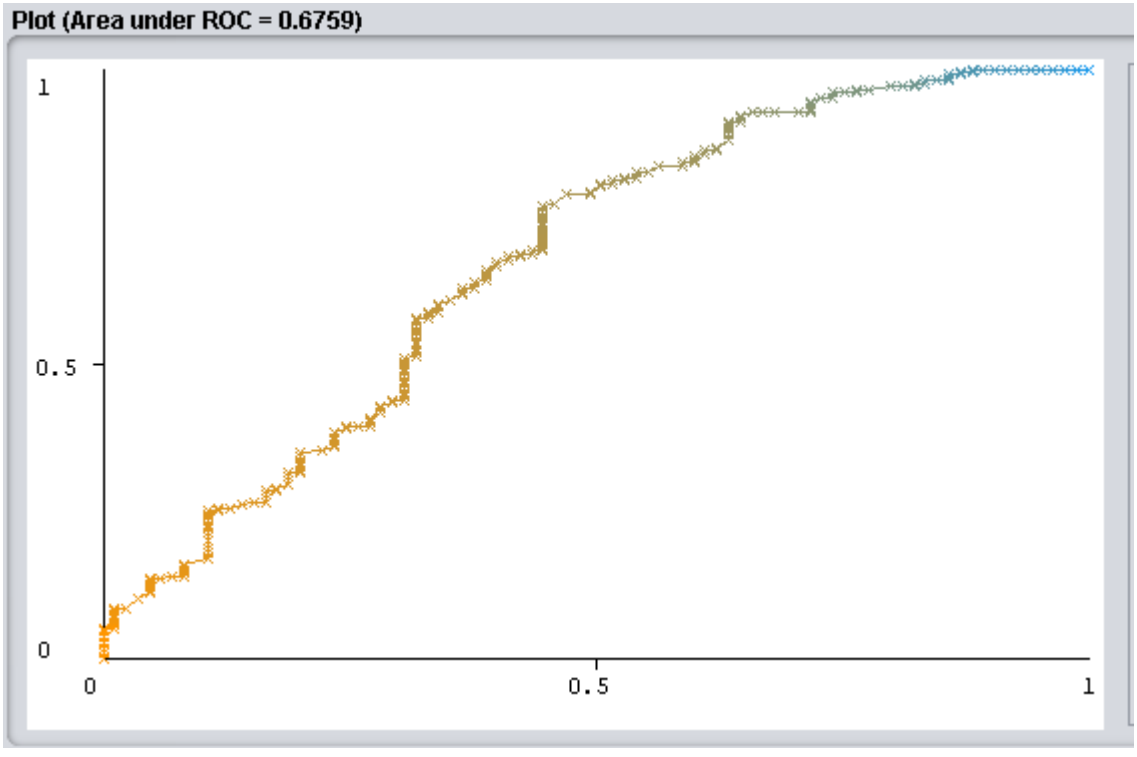
➔ ROC curve C4



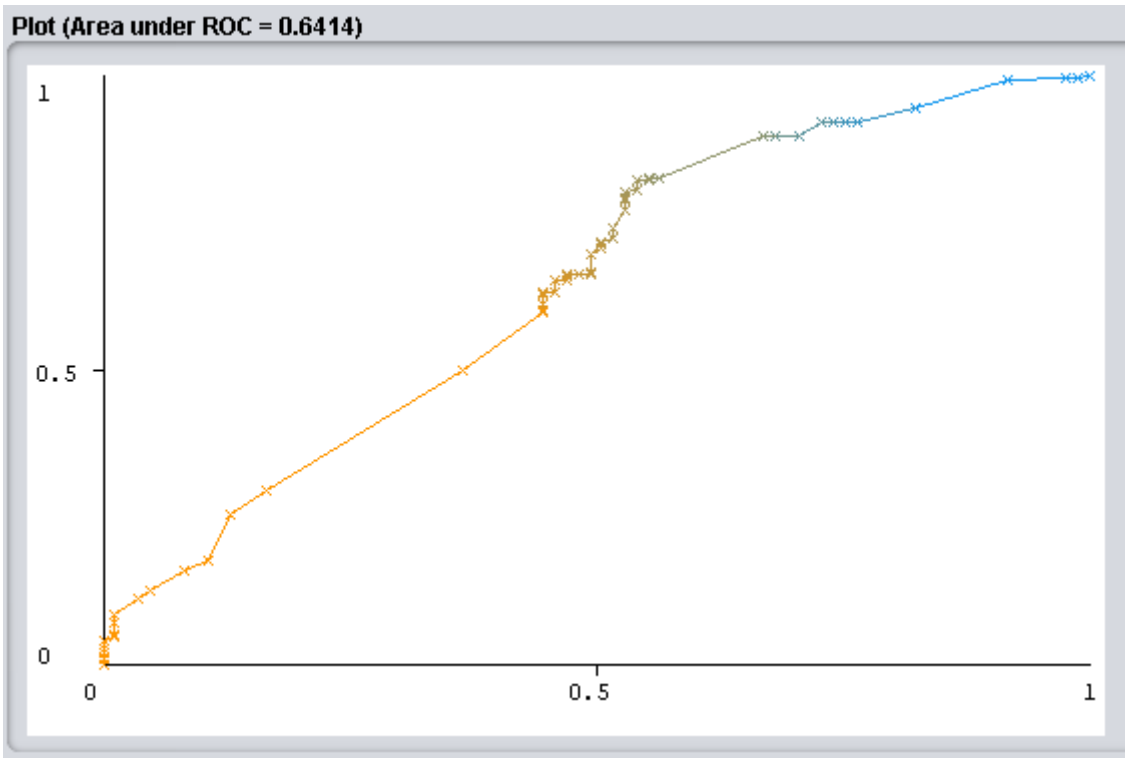
➔ ROC curve C5



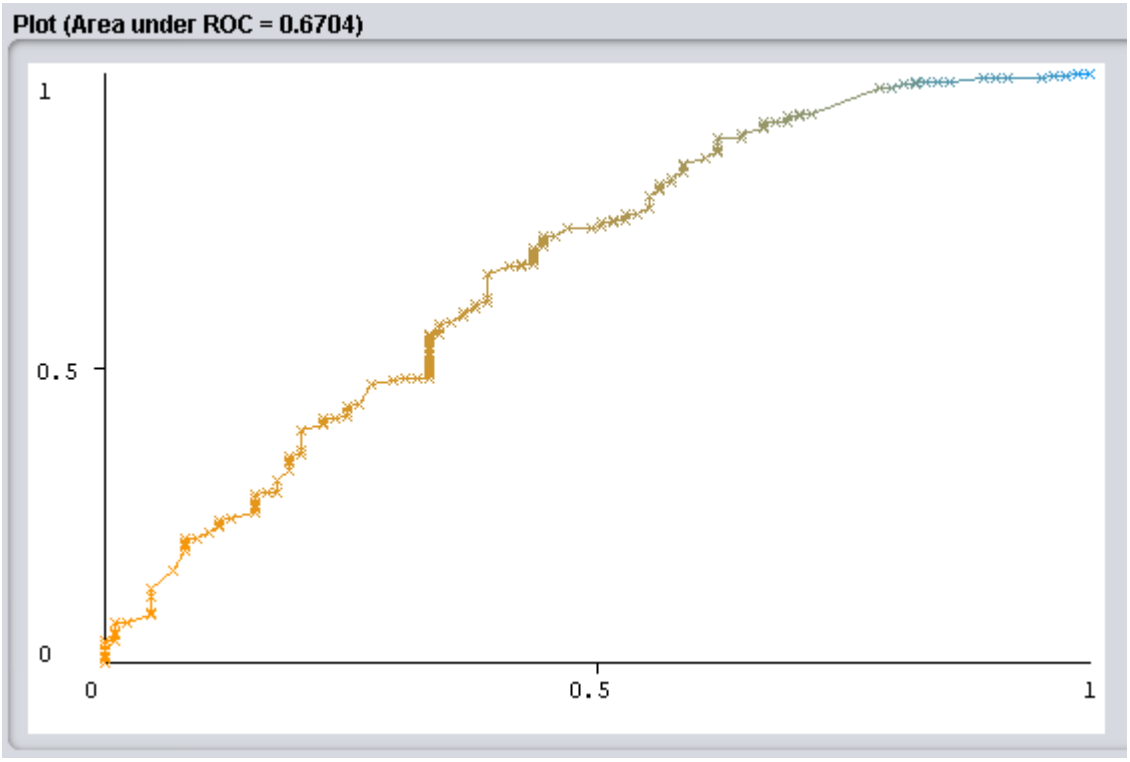
➔ ROC curve C6



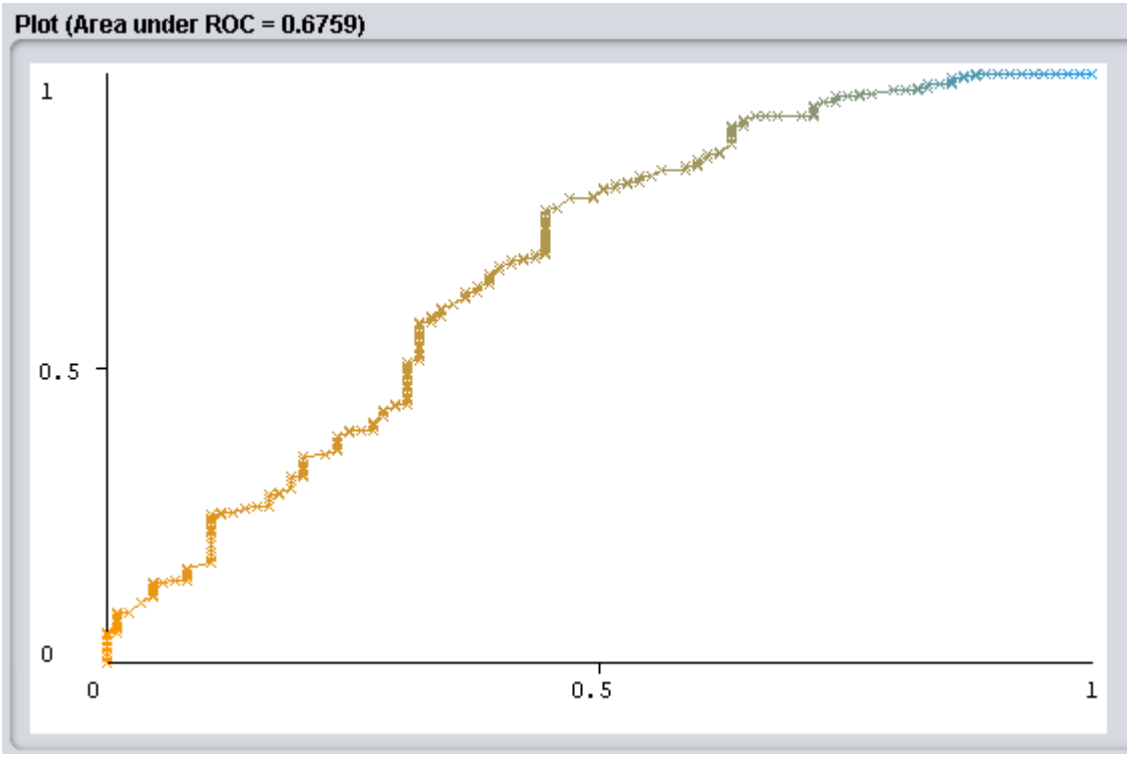
➔ ROC curve C7



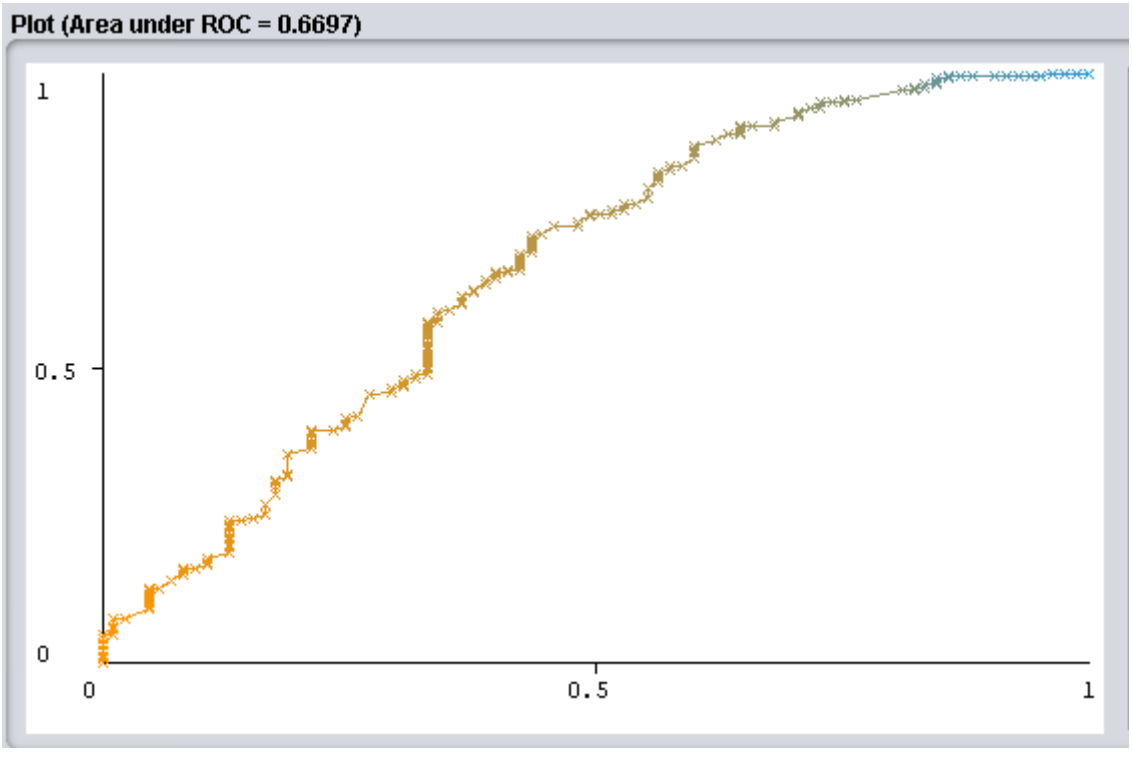
➔ ROC curve C8



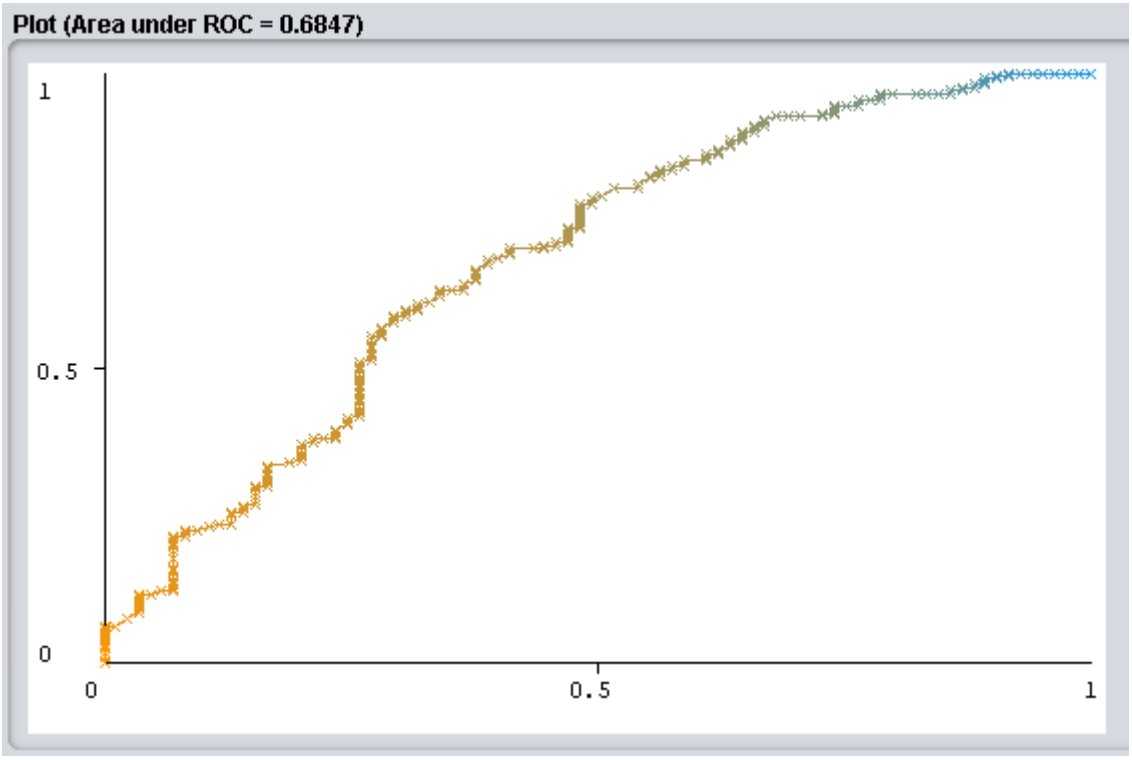
➔ ROC curve C9



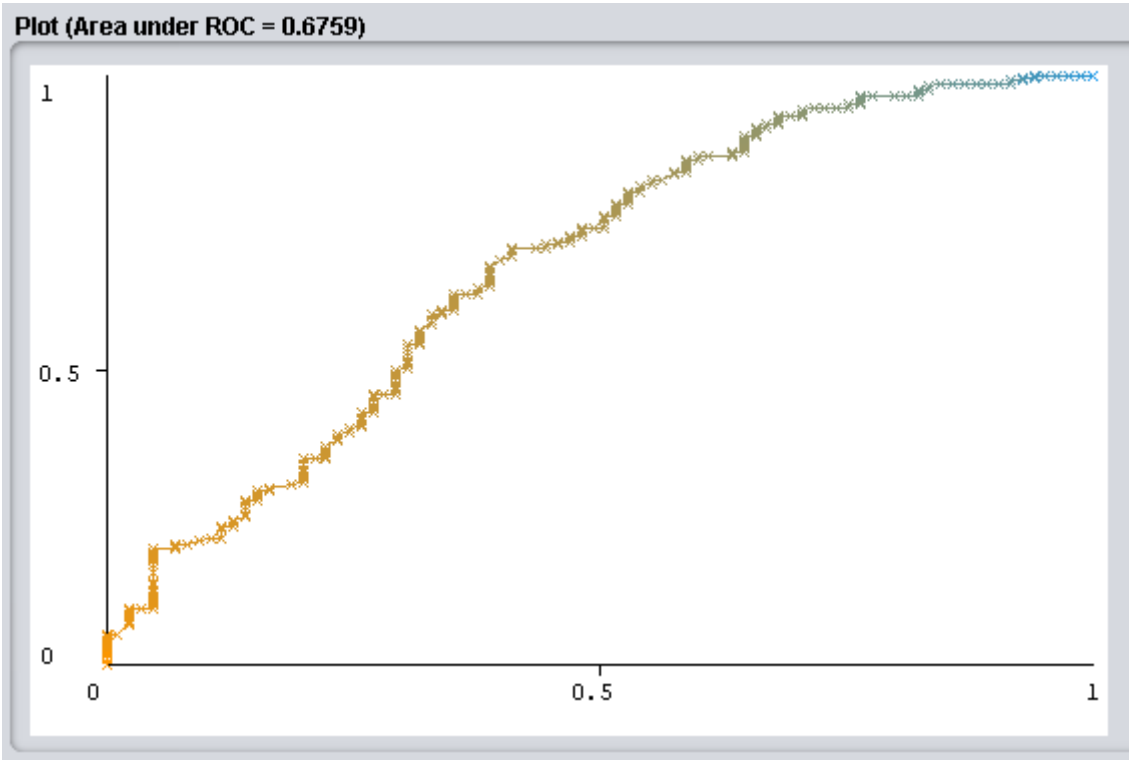
➔ ROC curve C10



➔ ROC curve C11



➔ ROC curve C12



HEPATITIS DATASET

➔ Introduction

The main idea of this classification problem is given a dataset about patients which suffer from hepatitis and they a series of characteristics which determine if they died or live because train a IBk classification model and evaluate its performance.

In this dataset we have 155 instances in total of which 129 of one class and 26 instances of another class. Each instance has 20 attributes, one of them is the class attribute. Because of we have only two classes we are faced with a binary classification problem.

➔ Number of classes

In this case our class attribute can take only ***two values***, *die* and *live*. Because of that we are faced to a ***binary classification problem***.

➔ Number of attributes

In this case we have **twenty attributes** which are *age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime, histology* and *class.*

➔ Number of samples

In this case in particular the number of samples or instances is **155.**

➔ **Performance metric values by configuration**

Configurations

Configuration	KNN	Distance Function
C1	1	Euclidean
C2	4	Euclidean
C3	9	Euclidean
C4	1	Manhattan
C5	4	Manhattan
C6	9	Manhattan
C7	1	Minkowski
C8	4	Minkowski
C9	9	Minkowski
C10	6	Minkowski
C11	12	Minkowski
C12	15	Minkowski

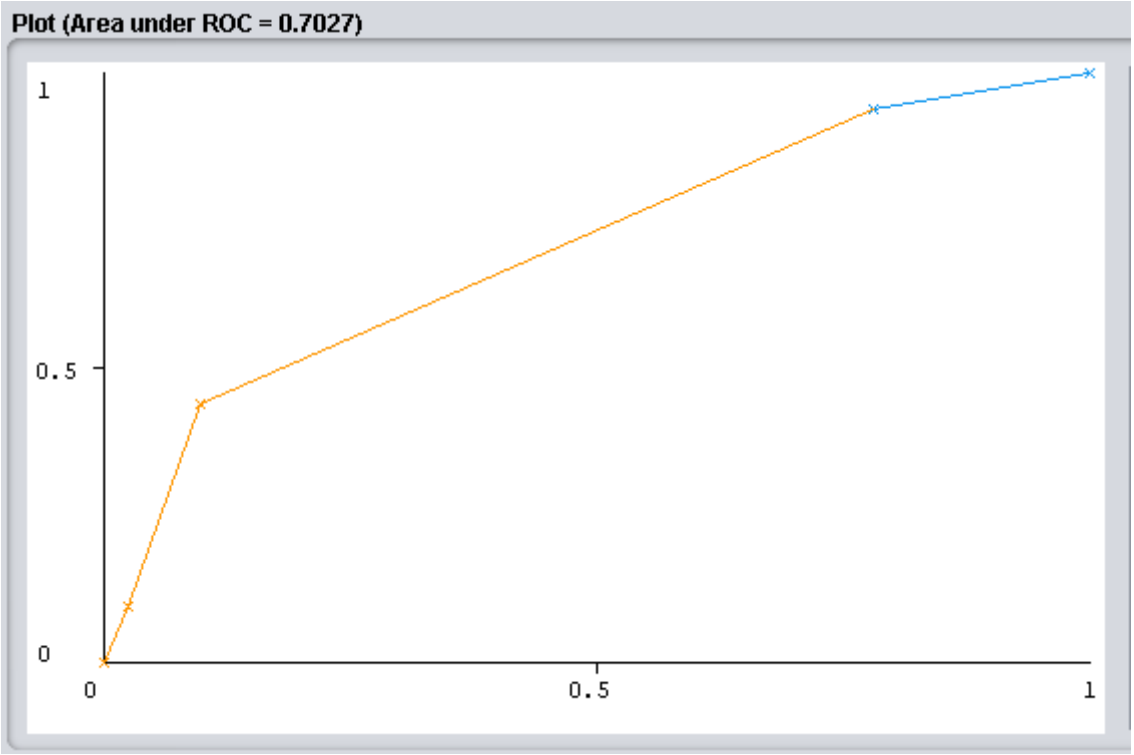
Results

Configuration	Correctly Classified%	TP rate	FP rate	Precision	Recall	Fmeasure	ROC
C1	80,6452	0,438	0,098	0,538	0,438	0,483	0,703
C2	82,5806	0,563	0,106	0,581	0,563	0,571	0,789
C3	85,1613	0,500	0,057	0,689	0,500	0,582	0,829
C4	81,2903	0,469	0,098	0,556	0,469	0,508	0,719
C5	82,5806	0,594	0,114	0,576	0,594	0,585	0,801
C6	85,1613	0,500	0,057	0,696	0,500	0,582	0,819
C7	80,6452	0,438	0,098	0,538	0,438	0,483	0,703
C8	82,5806	0,563	0,106	0,581	0,563	0,571	0,789
C9	85,1613	0,500	0,057	0,696	0,500	0,582	0,829
C10	85,1613	0,594	0,081	0,655	0,594	0,623	0,804
C11	82,5806	0,469	0,081	0,600	0,469	0,526	0,816
C12	81,2903	0,250	0,041	0,615	0,250	0,356	0,825

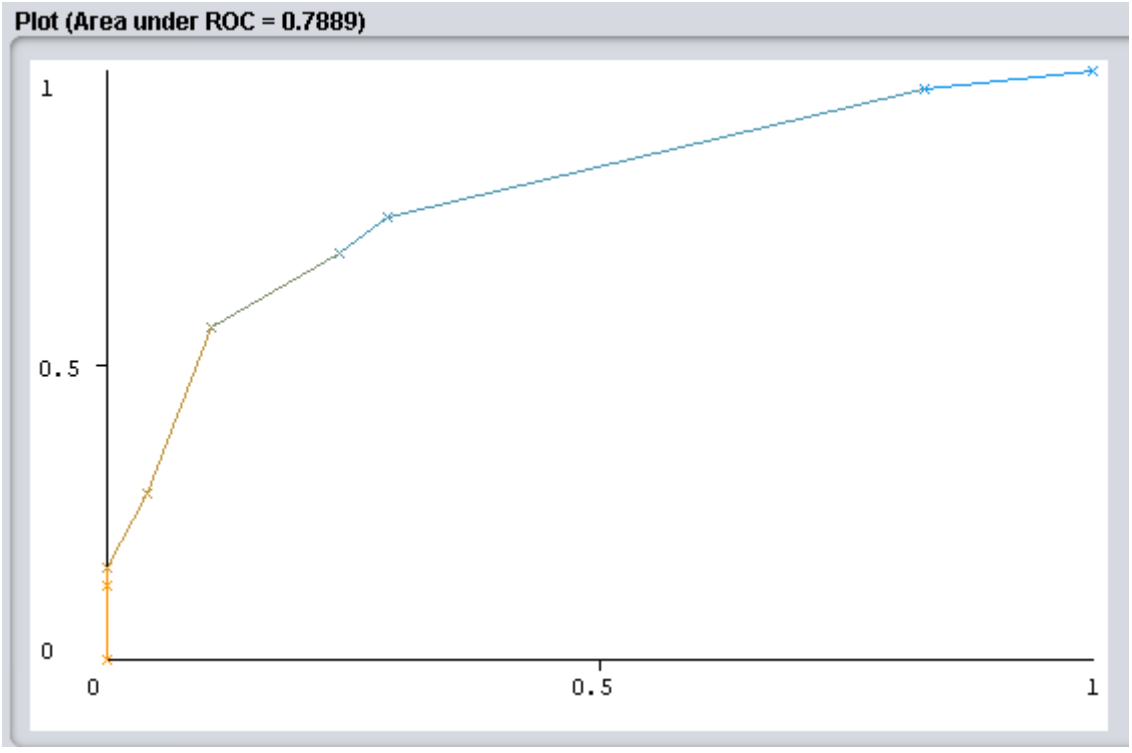
Section conclusions

As we can see again, configuration 1 is the one that gives the worst result, stating that a value of k = 1 creates a lot of noise in the classification. The best results in correctly classified instances can be found with k = 9 and without difference between distance functions, which give the same results at the same values of K. By way of conclusion we can say that this dataset behaves in a similarly in the tested configurations although with a notable improvement between the worst and the best result in ROC curve, which means that some instances create a small noise that can be eliminated by configuring the model.

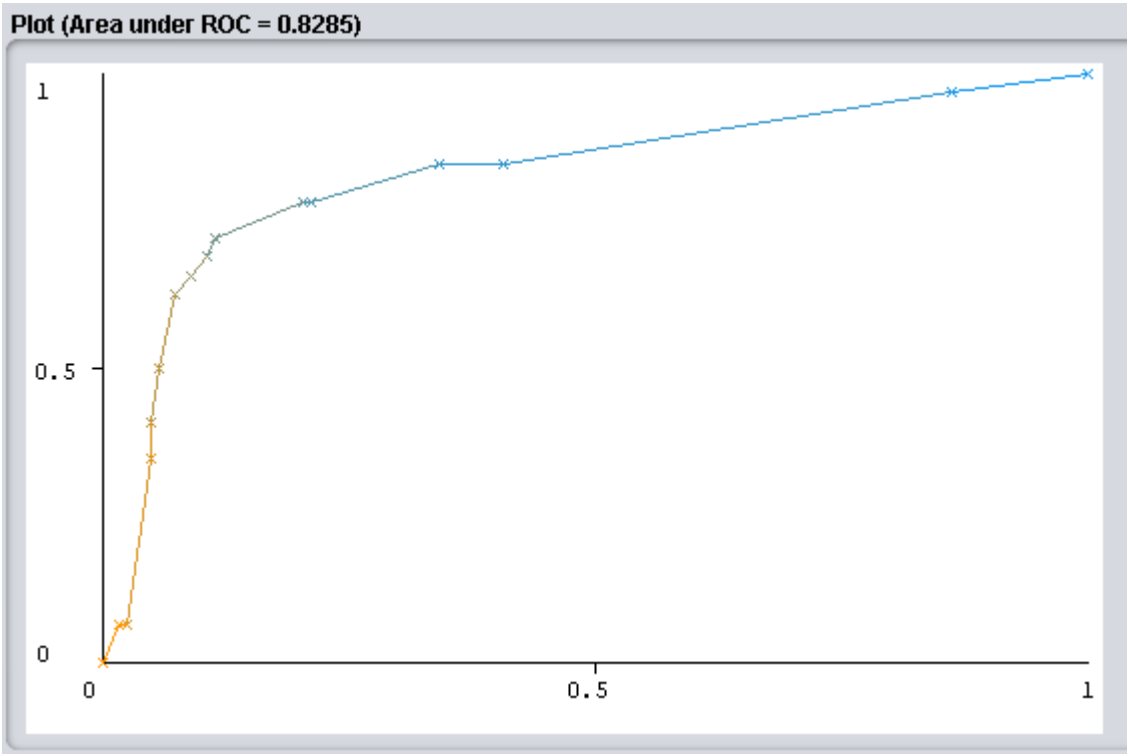
➔ **ROC curve C1**



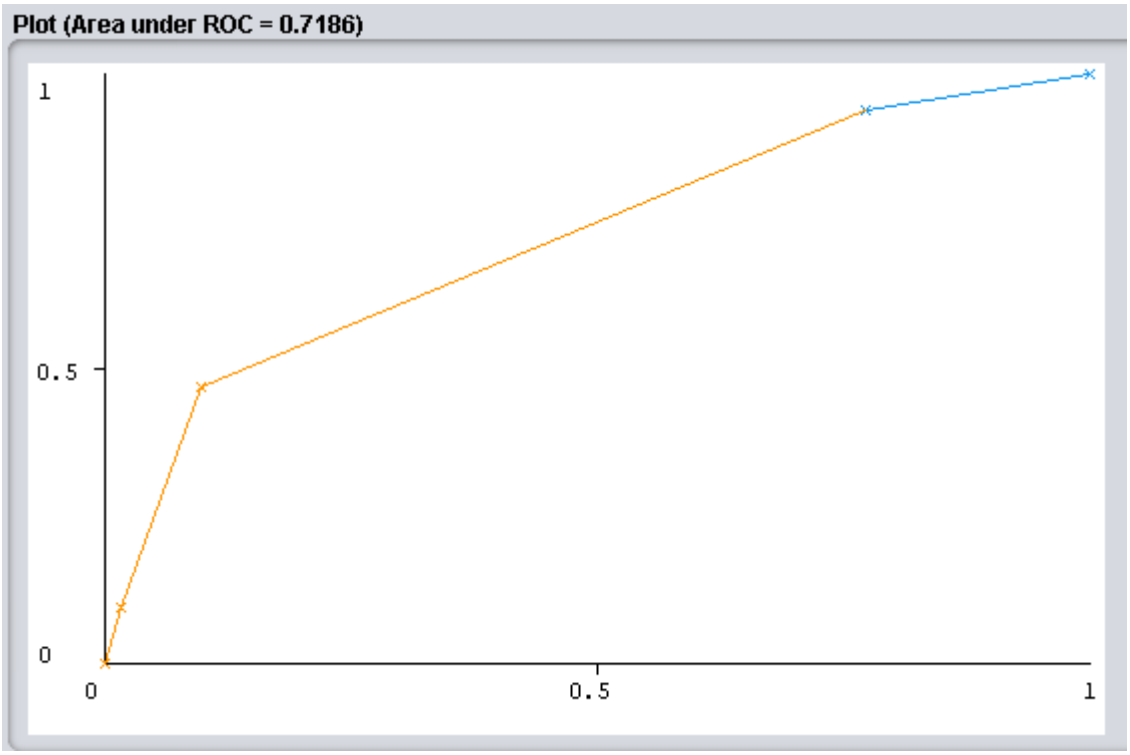
➔ **ROC curve C2**



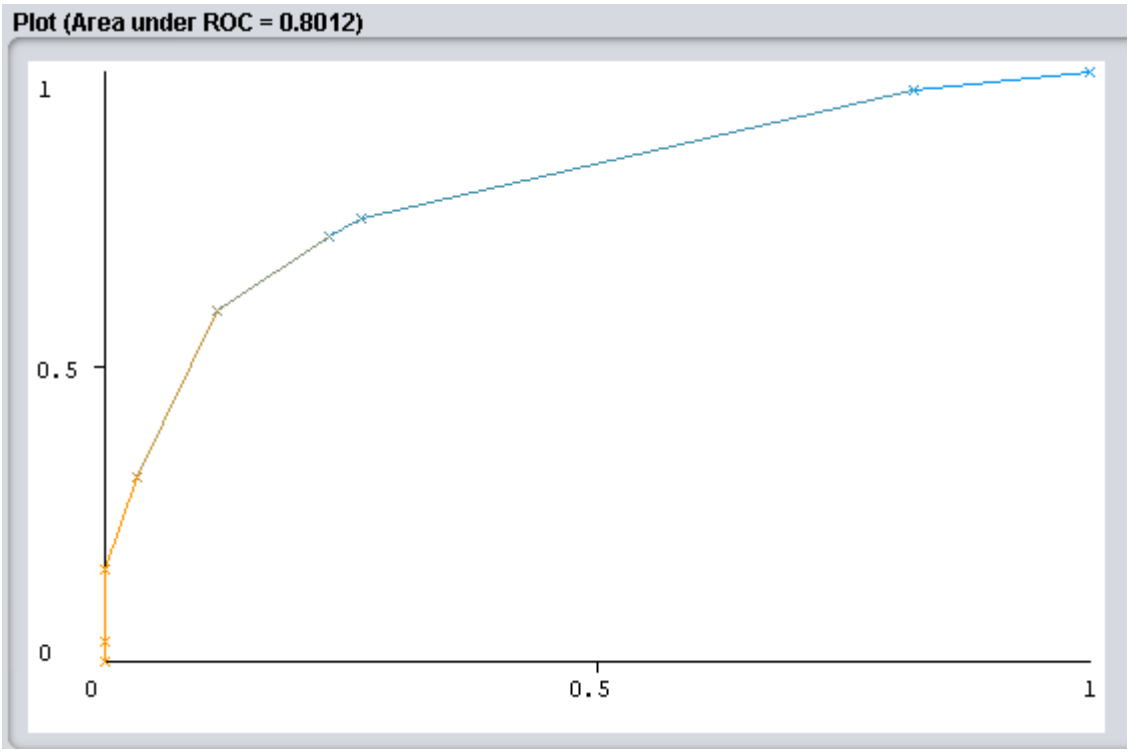
➔ ROC curve C3



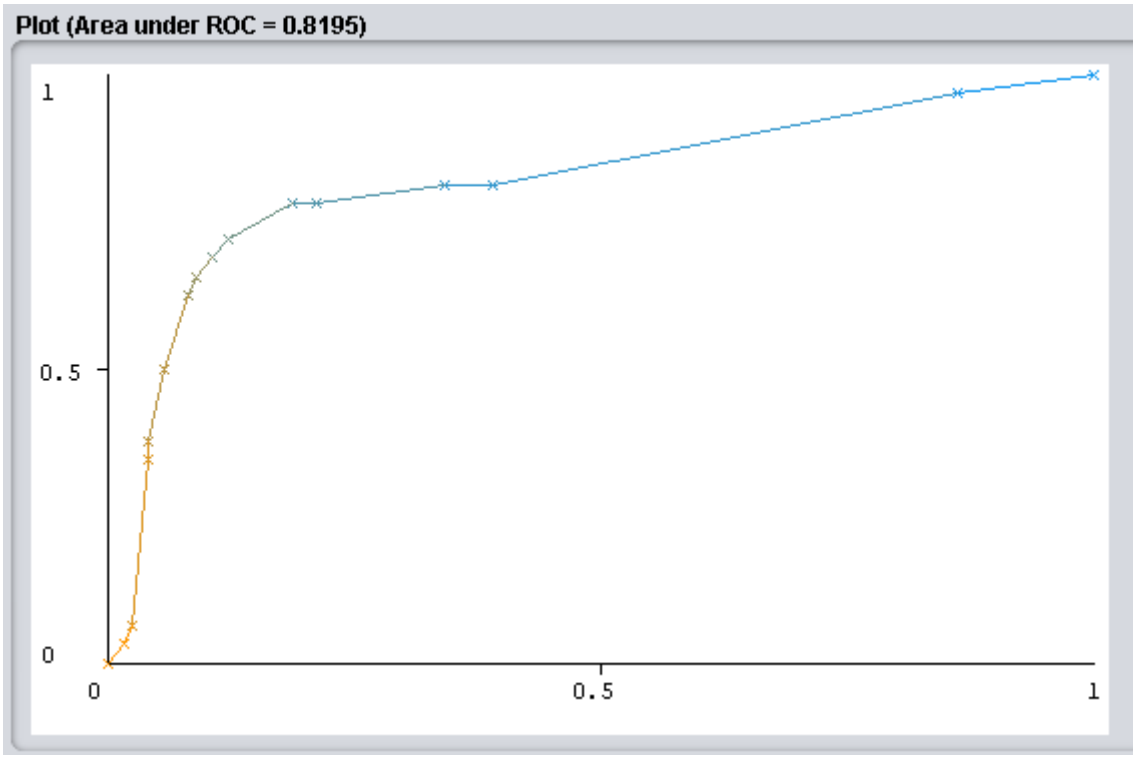
➔ ROC curve C4



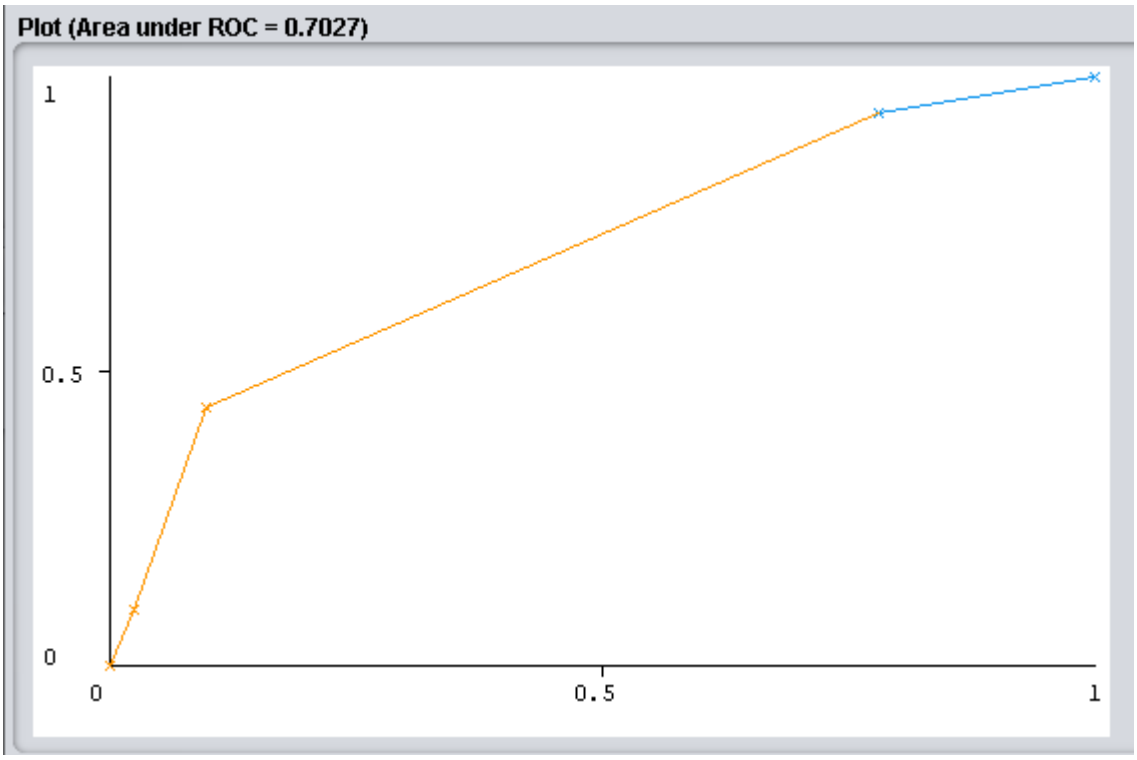
➔ ROC curve C5



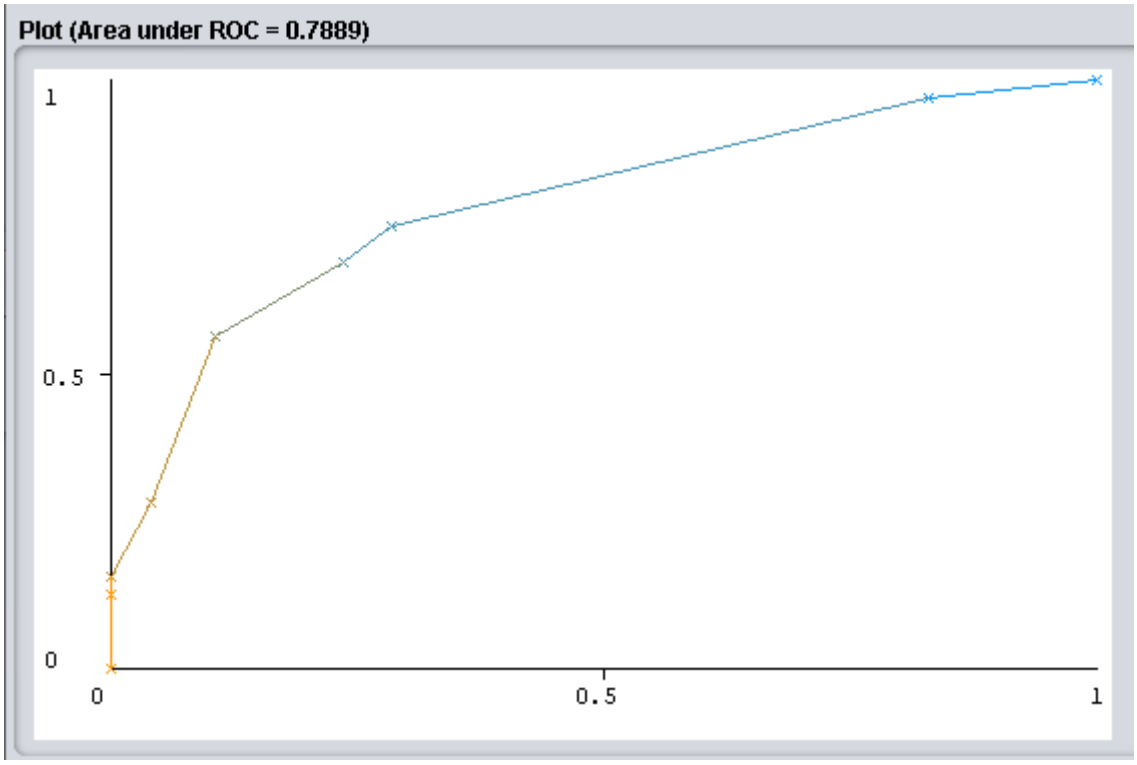
➔ ROC curve C6



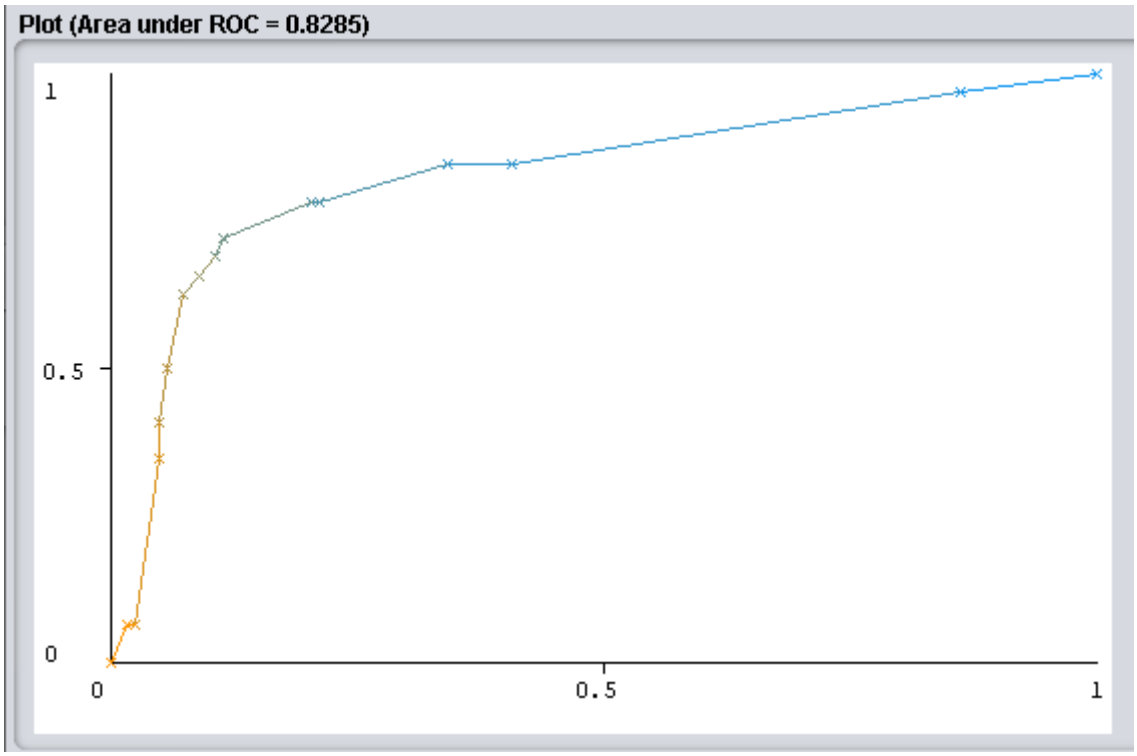
➔ **ROC curve C7**



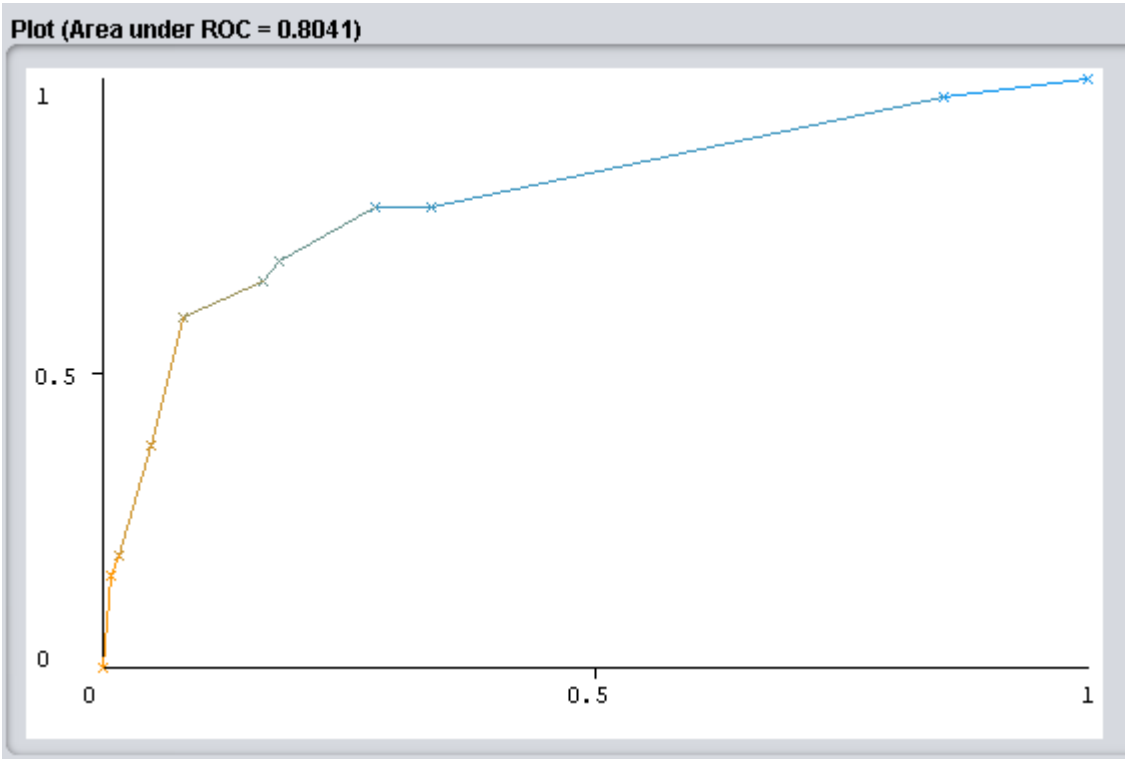
➔ **ROC curve C8**



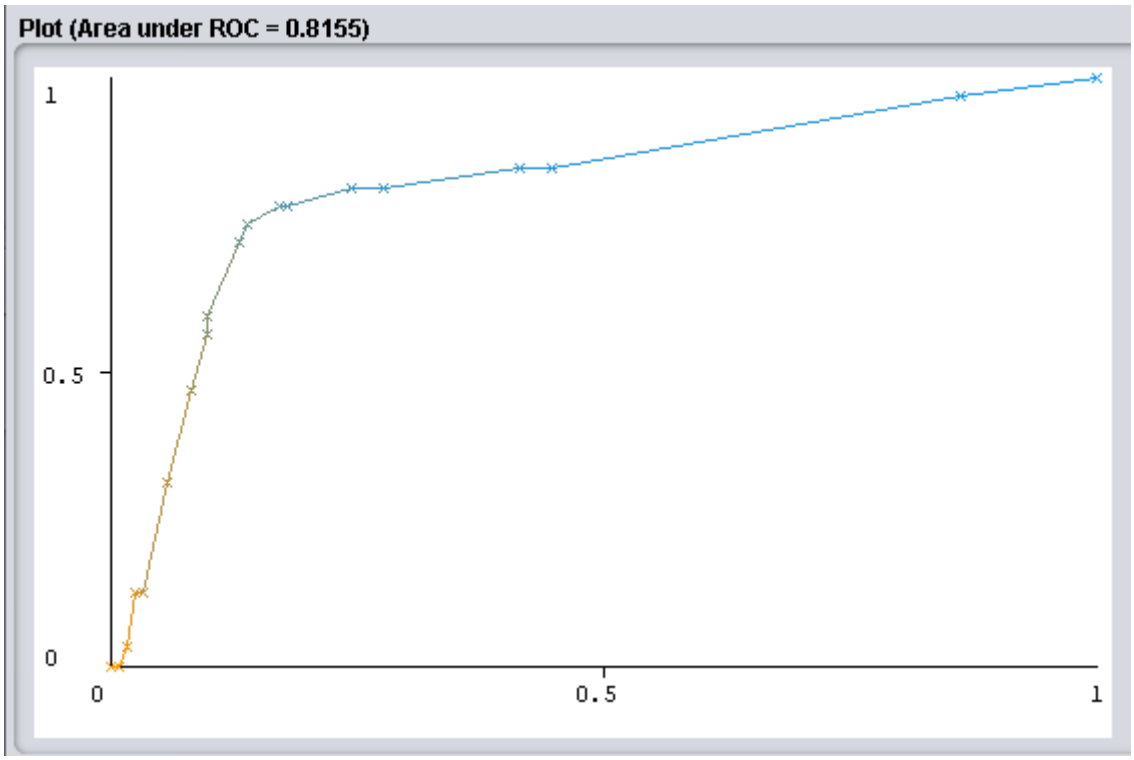
➔ **ROC curve C9**



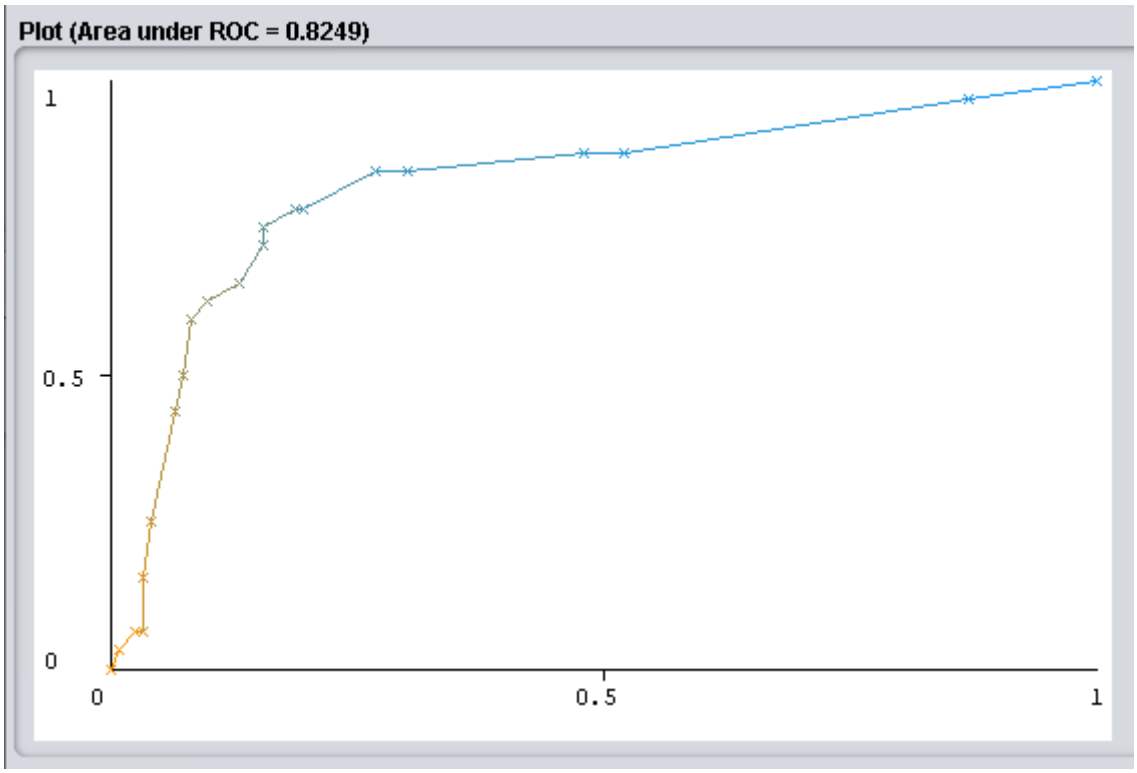
➔ **ROC curve C10**



➔ ROC curve C11



➔ ROC curve C12



POST-OPERATIVE PATIENTS DATASET

➔ Introduction

The main idea of this classification problem is given a dataset about patients which have passed an operation and they are in a postoperative recovery area waiting to a decision of where they should be sent next. Depending on a number of temperature measurements we train a IBk classification model and evaluate its performance.

In this dataset we have 90 instances in total of which they are divided in three categories or classes, 2 of class I, 24 of class S and 64 of class A. Each instance has 9 attributes, one of them is the class attribute. Because of we have three possible classes we are faced with a multiclass classification problem.

➔ Number of classes

In this case our class attribute can take only **three *values***, I, S and *A*. Because of that we are faced to a **multiclass *classification problem***.

➔ Number of attributes

In this case we have **nine attributes** which *l-core, l-surf, l-02, l-bp, surf-stbl, core-stbl, bp-stbl, comfort* and *class.*

➔ Number of samples

In this case in particular the number of samples or instances is **90.**

➔ **Performance metric values by configuration**

Configurations

Configuration	KNN	Distance Function
C1	1	Euclidean
C2	4	Euclidean
C3	9	Euclidean
C4	1	Manhattan
C5	4	Manhattan
C6	9	Manhattan
C7	1	Minkowski
C8	4	Minkowski
C9	9	Minkowski
C10	6	Minkowski
C11	12	Minkowski
C12	15	Minkowski

Results

Configuration	Correctly Classified%	TP rate	FP rate	Precision	Recall	Fmeasure	ROC
C1	66,666	0,667	0,704	?	0,667	?	0,414
C2	71,111	0,711	0,711	?	0,711	?	0,242
C3	71,111	0,711	0,711	?	0,711	?	0,385
C4	66,666	0,667	0,704	?	0,667	?	0,414
C5	71,111	0,711	0,711	?	0,711	?	0,242
C6	71,111	0,711	0,711	?	0,711	?	0,385
C7	66,666	0,667	0,704	?	0,667	?	0,414
C8	71,111	0,711	0,711	?	0,711	?	0,242
C9	71,111	0,711	0,711	?	0,711	?	0,385
C10	71,111	0,711	0,711	?	0,711	?	0,322
C11	71,111	0,711	0,711	?	0,711	?	0,378
C12	71,111	0,711	0,711	?	0,711	?	0,410

Section conclusions

In this dataset we have that the instances with the best results in correctly classified instances do not correspond to the best results in the ROC area. In relation to the configurations with a higher percentage of correctly classified instances we have a tie in all of which are greater than K = 1 and in ROC area we have that the Mankowski configurations with K = 1 and K = 15 are the most optimal.

➔ **Evaluation of the results**

Dataset	Best Result CorrectlyClassified	Best Result ROC	Best Result Precision
BreastCancer	C2,C5,C8,	C11	C12
Hepatitis	C3,C6,C9,C10	C3,C9	C3,C6,C9
Post-Operative	Tie K>1	C1,C4,C7	?

➔ **Differences between datasets**

Dataset	NumSamples	ClassType	ClassDistribution	Atributte Characteristics
BreastCancer	286	Binary	201/85	Linear/Nominal
Hepatitis	155	Binary	32/123	Nominal/ Numerical
Post-Operative	70	Multiclass	2/24/64	Nominal/ Numerical

➔ **Last conclusions**

By way of conclusion we can say that the choice of the number of closest neighbors is a fundamental parameter in the results. Introducing ourselves more in this aspect we have verified that k = 1 has been one of the worst results that we have obtained, verifying that for a correct classification we must normally take into account more than one point, on the other hand an excessively high number does not correspond to a better result causing overfiting and being able to classify based only on the proportions of classes in our dataset. Analyzing the global results of the three datasets we have that the best results in correctly classified instances have been given in configurations with k = 4 for BreastCancer or K = 9 for Hepatitis, in the case of PostOperative it provides us with the same results with values of K> 1. In the ROC curve results we have that the values of K = 12 and K = 9 give us the best results in BreastCancer and Hepatitis respectively, in PostOperative curiously the best value is K = 1. Finally we can say that KNN has been an easy algorithm to understand and configure, K being the only hyperparameter to modify, in turn it would also highlight the limitations it has due to its complexity in time and space for datasets with high dimensionality. Changes in the distance function, normalizing the data or reducing the dimensionality are ways to reduce the inconveniences that this presents.